International Journal of Forecasting 8 (1992) 559–573
North-Holland

The effects of feedback and training on the performance of probability forecasters

P. George Benson
Carlson School of Management, University of Minnesota, Minneapolis, MN 55455, USA

Dilek Önkal
Department of Management, Bilkent University, 06533 Ankara, Turkey

Abstract: An experiment examined the effects of outcome feedback and three types of performance feedback – calibration feedback, resolution feedback, and covariance feedback – on various aspects of the performance of probability forecasters. Subjects made 55 forecasts in each of four sessions, receiving feedback prior to making their forecasts in each of the last three sessions. The provision of calibration feedback was effective in improving both the calibration and overforecasting of probability forecasters, but the improvement was not gradual; it occurred in one step, between the second and third sessions. Simple outcome feedback had very little effect on forecasting performance. Neither resolution nor covariance feedback affected forecasters' performances much differently than outcome feedback. However, unlike outcome feedback, the provision of performance feedback caused subjects to manage their use of the probability scale. Subjects switched from two-digit probabilities to one-digit probabilities, and those receiving calibration and resolution feedback also reduced the number of different probabilities they used.

Keywords: Probability forecasting, Judgmental forecasting, Subjective probability, Outcome feedback, Performance feedback, Scoring rules, Calibration, Resolution, Covariance decomposition.

1. Introduction

In this paper we investigate the effects of different forms of performance feedback and associated training on the quality of judgmental probability forecasts provided by individual forecasters. Performance feedback is one of four types of feedback that are relevant in judgmental forecasting tasks; the others are outcome, process, and environmental feedback. All are defined below:

- *Outcome feedback* is information about the realization of a previously predicted event.
- *Performance feedback* is information about the accuracy of the forecaster's predictions. It is derived from the forecaster's predictions and the outcomes that occur.
- *Process feedback* is information about the forecaster's cognitive processes. It includes information about the evidence perceived by the forecaster, how the forecaster utilizes evidence in developing predictions, and information about the predictions themselves.
- *Environmental feedback* (or task feedback) is information about the event to be predicted, including the factors that may influence the event and their relationship to the event.

[For a more general description of these feedback types, see Balzer et al. (1989).] Feedback research in the area of probability forecasting

Correspondence to: P.G. Benson, Carlson School of Management, University of Minnesota, 271 19th Avenue South, Minneapolis, MN 55455, USA.

has been concerned with outcome and performance feedback. We know of no probability forecasting studies that have addressed process or environmental feedback.

Outcome feedback has proven to be ineffective for improving the accuracy of probability forecasts [Fischer (1982)]. It provides neither the information forecasters need to understand the key relationships in the environment [Brehmer (1980)], nor the information that forecasters would find useful for calibrating their probability forecasts to better reflect the relative frequency of occurrence of the forecasted events. Consider a securities analyst who is responsible for forecasting the direction of change in the quarterly earnings of a particular firm. Knowledge of only the firm's actual quarterly earnings – outcome feedback – clearly is not sufficient for the analyst to either understand the forces that caused those earnings or know how to improve her usage of the probability scale in developing the next quarter's forecast.

Two kinds of performance feedback have been investigated: scoring-rule feedback and calibration feedback. A scoring rule assigns a reward or penalty to a forecaster as a function of the forecaster's reported probability forecasts and the outcomes that occur [Winkler (1969), Friedman (1983)]. Scoring rules are typically designed to indicate the extent of a forecaster's 'external correspondence' – i.e. the extent to which the forecaster assigns probabilities close to 1 for events that occur and probabilities close to 0 for events that do not occur [Yates (1982)]. Scoring-rule feedback consists of the forecaster's score for a set of probability forecasts.

Laboratory experiments have yielded mixed results for scoring-rule feedback. Staël von Holstein (1972) and Fischer (1982) concluded that such feedback had no effect on the forecasting performances of their subjects. However, based on an experiment similar to Staël von Holstein's, Kidd (1973) [cited in Beach (1975)] concluded that scoring-rule feedback was effective and could be used to improve the accuracy of probability forecasters.

Calibration feedback provides forecasters with information about their ability to assign appropriate probabilities to outcomes. A forecaster is said to be well-calibrated if for all predicted outcomes assigned a given probability, the proportion of those outcomes that occur (referred to below as the 'proportion correct') is equal to the probability. For example, if it actually rained on 40% of the days that a weather forecaster predicts a 0.4 chance of rain, the forecaster's 0.4 probability forecasts are well-calibrated. If a variety of the weather forecaster's other probability forecasts similarly match event frequencies, the weather forecaster is well-calibrated. The calibration component of the Brier Scoring Rule can be used to evaluate calibration [Brier (1950), Murphy (1973)]. Calibration feedback has not been standardized. At a minimum, it consists of numerical summaries and/or graphical displays of the reported probabilities, the proportion correct associated with each probability value, and the number of assessments of each value.

Calibration feedback appears to be a promising means of improving the performance of probability forecasters. Both individualized and group feedback have proven to be effective in field studies of weather forecasters even though only one feedback session was employed [Murphy and Daan (1984), Murphy et al. (1985)].

Except for scoring-rule feedback, performance feedback in forecasting tasks has not been evaluated in the laboratory. In this paper we report the results of a laboratory experiment that investigated the effects of three different forms of performance feedback – calibration feedback, resolution feedback, and covariance feedback – on the performance of probability forecasters.

Section 2 briefly reviews two feedback studies that are relevant for motivating the current study and for understanding and interpreting its results. Section 3 describes the experiment. Sections 4 and 5 present and discuss the results of the experiment, respectively. Section 6 presents conclusions and directions for future research.

2. Relevant performance feedback studies

Studies of the effects of performance feedback on subjective probability assessments fall into two categories: those concerned with probability forecasts and those concerned with assessments of confidence in answers to general-knowledge questions. In the latter case, subjects are typically asked to answer almanac-type questions (e.g. Is the population of Minnesota greater than the

population of Wisconsin?) and to provide subjective probabilities that reflect their confidence in their answers. Such 'general-knowledge tasks' have received considerably more attention from researchers than probability-forecasting tasks. [For a review of this literature, see Lichtenstein et al. (1982).]

Even though the focus of the present paper is forecasting, there are several reasons for being concerned with previous studies of general-knowledge tasks. First, there is the possibility that results from such studies can be generalized to forecasting tasks. Fischhoff and MacGregor (1982) argue that judgments of confidence in answers to general-knowledge questions are similar to judgments of confidence in forecasts (i.e. probability forecasts). On the basis of their empirical results, they conclude that '. . .one should have considerably increased confidence in extrapolating the results of earlier [general-knowledge-task] research to confidence in forecasts' (p. 169). However, Wright and Ayton (1986) and Ronis and Yates (1987) argue to the contrary. We come down on the side of the latter two papers, as will be explained in Section 5. Second, the general-knowledge studies of Lichtenstein and Fischhoff (1980) and Sharp et al. (1988) have informed the design of the present study. Third, Lichtenstein and Fischhoff's empirical results serve as a basis of comparison for the results of the present paper. Accordingly, we briefly describe pertinent aspects of these two studies.

A consistent finding of general-knowledge studies is that the calibration of subjects' confidence judgments is at best fair, with the predominant reason for miscalibration being overconfidence. Subjects apparently believe that they have greater knowledge than they actually possess [Fischhoff et al. (1977)]. For example, of the answers in which subjects are totally confident and so indicate with a probability of 1.0, only 85% may be correct. Of the answers assigned a probability of 0.8, only 60% may be correct.

Lichtenstein and Fischhoff (1980) investigated the use of calibration feedback and associated training as a means of eliminating such miscalibration. In each of 11 sessions, 12 subjects answered 200 two-alternative, general-knowledge questions. For each question, subjects chose the answer they believed to be correct and assigned a probability between 0.5 and 1.0 to indicate their confidence in the chosen answer. After each of the 11 sessions, subjects received individualized performance feedback that included calibration feedback, a measure of their overconfidence (described later), their Brier Score, and the knowledge, calibration, and resolution components of their Brier Score derived using Murphy's (1973) decomposition (i.e. Brier Score = Knowledge Score + Calibration Score − Resolution Score).

Lichtenstein and Fischhoff found that feedback significantly improved subjects' calibration, and that virtually all of the improvement came between the first and second feedback sessions. In a second experiment, using only three feedback sessions, the same results were obtained.

Lichtenstein and Fischhoff also evaluated the resolution performance of their subjects. In the context of general-knowledge studies, resolution refers to an individual's ability to discriminate between answers that are correct and incorrect by differentially assigning probabilities to correct and incorrect answers. Lichtenstein and Fischhoff evaluated their subjects' resolution using the resolution-component of the Brier Scoring Rule. In both experiments, resolution was basically unaffected by feedback.

Sharp et al. (1988) also investigated the effects of performance feedback on confidence judgments. In each of four sessions one week apart, 54 subjects (of which 27 comprised a feedback group and 27 a control group) answered 60 general-knowledge questions and reported subjective probabilities that represented their confidence in the chosen answers. At the start of sessions 2, 3, and 4, subjects were given calibration feedback from the previous session. In addition, each subject was given the average of the probabilities she assigned to correct answers and the average probability for incorrect answers.

Unlike Lichtenstein and Fischhoff (1980), they found that feedback did not significantly influence calibration or overconfidence, but that it did influence subjects' resolution. The feedback group's resolution performance improved across the four sessions relative to the control group's. What makes this result particularly interesting is that '. . .the feedback contained no strategic information which would lead to im-

proved resolution' [Sharp et al. (1988, p. 280)].

The difference in the resolution results of the two studies can in part be explained by the different resolution measures used in the studies. Sharp et al. argue that the resolution component of the Brier Scoring Rule, which was the measure used by Lichtenstein and Fischhoff, is inappropriate. They point out that (1) it is bounded above by the knowledge score of the Brier-Score decomposition, and (2) knowledge scores differ among subjects. As a result, they maintain that resolution performance should not be evaluated by comparing subjects' raw resolution scores, but by computing each subject's resolution-score to knowledge-score ratio (called η^2) and comparing them. We employ this resolution measure in our analysis.

The present study investigates the effects of calibration and resolution feedback on probability forecasters in a laboratory setting. In addition, the effects of covariance feedback (described in the next section), another form of performance feedback, are studied. These three types of performance feedback are compared with simple outcome feedback. We expected each type of performance feedback to improve forecasting performance. In particular, we expected calibration feedback to improve forecasters' calibration scores; resolution feedback to improve forecasters' resolution scores; and covariance feedback to improve forecasters' slope and scatter scores (described later) of the covariance decomposition of their Brier Scores. We did not expect outcome feedback to improve forecasting performance.

3. Method

3.1. Subjects and task

Eighty paid subjects from the University of Minnesota who expressed an interest in college football began the four-week-long experiment. Each subject was randomly assigned to one of three feedback groups – calibration feedback, resolution feedback, or covariance feedback – or to a control group. Owing primarily to the length and time commitment required by the experiment, a number of subjects dropped out of the study while it was in progress. Fifty-two subjects (41 undergraduate students, 10 graduate students, and one faculty member) completed the experiment. The calibration, resolution, and covariance feedback groups were comprised of 15, 11, and 16 subjects, respectively; 10 subjects served in the control group.

The experiment involved four weekly forecasting sessions. In each session, subjects were asked to make probability forecasts for the outcomes of 55 major college football games that were to be played the following weekend. Subjects were given a list of the games to be predicted (with home and visiting teams identified) one week prior to the forecasting session. For each of the games in each session, subjects were asked to predict the winning team and to assess a subjective probability (between 0.5 and 1.0, inclusive) that reflected that team's chances of winning. This is referred to as a two-alternative, half-range, assessment task [Lichtenstein et al. (1982)]. Recent evidence suggests that such assessment structures may be superior to full-range tasks (i.e. asking for a probability between 0 and 1.0) in forecasting problems [Ronis and Yates (1987)]. At the beginning of each of the last three sessions, subjects in the feedback groups received performance feedback derived from their predictions from the previous weeks. Control group subjects received only outcome feedback.

3.2. Training and feedback

At the beginning of the first session all subjects received approximately 1 h of training in subjective probability and probability forecasting, including training in two-alternative, half-range, probability forecasting tasks. At the beginning of each of the remaining three sessions, each subject received a listing of her predictions and probability assessments from the previous week, along with the actual outcomes of the 55 games (i.e. game scores). All groups including the control group received this outcome feedback. Subjects in the treatment groups also received performance feedback.

The provision of outcome feedback to all groups seemed appropriate since game scores were available to the subjects outside the laboratory through newspapers, television, etc.

Furthermore, in most real forecasting tasks (e.g. weather forecasts, sales forecasts) forecasters receive outcome feedback. Also, it made it possible for the experimenter to interact with and motivate control group subjects.

At each session, all feedback from earlier sessions was available to the subjects. In all groups, subjects were encouraged to participate in one-on-one discussions with the experimenter concerning their personal feedback. Since the informational and motivational aspects of feedback and training were confounded, there was a deliberate attempt to provide similar attention and motivation to all groups and all subjects.

3.2.1. Control group

Control group subjects received (1) outcome feedback and (2) their ranking within the group as determined by their Brier Scores. Subjects were not informed of the criteria used to establish rankings. The ranking was used to motivate performance.

3.2.2. Calibration feedback group

Subjects in the calibration feedback group received (1) outcome feedback, (2) their individual calibration scores computed using the calibration component of Murphy's decomposition of the Brier Score, (3) their individual calibration curves (described below), (4) the best and the average calibration scores for their group, and (5) their ranking within the group as determined by calibration scores. Items (4) and (5) were used to motivate performance.

The computation of the calibration component requires that all probability forecasts be grouped into categories (0.50 to 0.59, 0.60 to 0.69, etc.). A calibration curve is constructed by plotting the proportion of correct forecasts in each category against the respective mean probability forecast for the category. The resulting points are then connected with line segments, beginning with the lowest mean probability forecast and proceeding to the highest [Lichtenstein et al. (1982)].

At the beginning of the second, third, and fourth sessions, the concept of calibration and the calculation of the calibration score were discussed in detail. Calibration curves were explained and perfect calibration, overforecasting, and underforecasting were illustrated through examples. A forecaster is said to be *overforecasting* (underforecasting) when her probability forecasts tend to be larger (smaller) than the proportion of correct forecasts [Murphy and Daan (1984)]. In studies of confidence judgments, this phenomenon is referred to as *overconfidence* (underconfidence). Subjects were encouraged to construct probability forecasts in a manner that would 'move' their calibration curves from earlier sessions as close to the 45° line (perfect calibration) as possible.

3.2.3. Resolution feedback group

Subjects in the resolution feedback group were given (1) outcome feedback, (2) their individual resolution scores computed using the resolution component of Murphy's decomposition of the Brier Score, (3) their individual calibration curves, (4) the best and the average resolution scores for the group, (5) their ranking within their group as determined by resolution scores, and (6) a knowledge-level analysis of their forecasts. Items (4) and (5) were used to motivate performance. As in Sharp et al.'s (1988) study, the feedback contained no new information about the events being forecasted.

The last feedback item was designed to help subjects attain maximally different proportions of correct forecasts for their reported probabilities and thereby improve their resolution scores. To develop this feedback item, all resolution-group subjects were asked to indicate at the time they made their predictions whether their knowledge of the game in questions was 'none', 'very little', 'a fair amount', or 'very extensive'. Subjects were specifically instructed to indicate their knowledge level prior to making a prediction. Examples of different knowledge levels were discussed with the subjects. Subjects were given feedback on the knowledge levels associated with each probability category they used. For each probability category, a tabular frequency distribution describing the knowledge levels was supplied.

At the beginning of the second, third, and fourth sessions, the concept and definition of resolution were discussed in detail. Calibration curves were explained and examples used to illustrate good vs. poor resolution. Subjects were encouraged to categorize their forecasts using probability classes whose proportions correct

would be maximally different from the overall proportion of correct forecasts. Subjects were urged to rely on their knowledge of each game in constructing probability forecasts. They were instructed to use the knowledge-level distributions to help them see the relationship between the probabilities they used and their level of knowledge. They were advised that high probabilities should be associated with high knowledge levels, but that low probabilities could be associated with either low or high knowledge levels. Thus, they were urged not to assess high probabilities unless they had both a high knowledge level and strong evidence favoring a particular team.

3.2.4. *Covariance feedback group*

The feedback supplied to the covariance feedback group was derived from the covariance decomposition of the Brier Score, as were certain measures used later in the paper to evaluate the performances of forecasters in all of the experimental groups. Accordingly, before proceeding with a description of covariance feedback, we briefly review relevant aspects of the covariance decomposition.

3.2.4.1. The covariance decomposition. Instead of focusing on groupings or categories of similar probabilities as does Murphy's decomposition, the analytic strategy of the covariance decomposition involves grouping probabilities according to whether or not they are associated with the occurrence of a target event [Yates (1982)]. For example, in weather forecasting, precipitation forecasts could be grouped according to whether it actually rained (the target event) or not on the days for which forecasts were made. Yates (1982) recommended that 'covariance graphs' be used for (1) interpreting the components of the decomposition and (2) uncovering systematic differences that may exist in the probability forecasts reported when the target event does and does not occur [see also Yates and Curley (1985)]. An example covariance graph is presented in Exhibit 1.

The covariance graph consists of two histograms that are conditional distributions of probability forecasts: one for when the target outcome occurred ($d = 1$) and one for when it did not ($d = 0$). The mean of the former distribution is indicated by $\bar{f}_1$; the mean of the latter by $\bar{f}_0$.

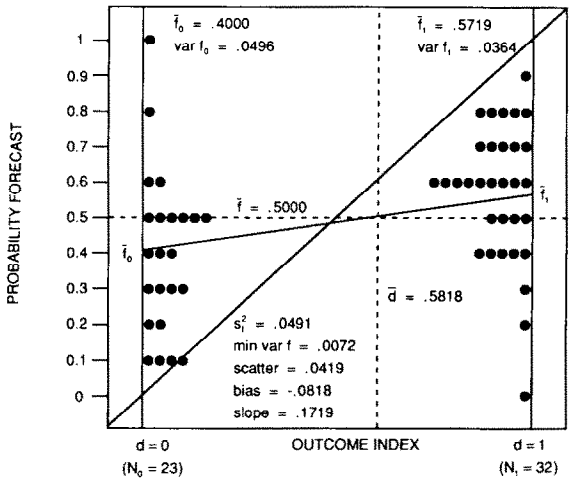

Exhibit 1. Examplar covariance graph.

The horizontal dotted line marks the overall mean probability forecast, $\bar{f}$. The variances of the two conditional distributions are denoted by var f_1 and var f_0, respectively. The overall variance of the reported forecasts is denoted by s_f^2 and is referred to as the 'forecast variance'. The vertical dotted line indicates the overall relative frequency of the target outcome's occurrence, $\bar{d}$.

In the remainder of this subsection we use the covariance graph to help describe three elements of the covariance decomposition that are used later in the paper: slope, scatter, and bias. Consider the line connecting the points $(0, \bar{f}_0)$ and $(1, \bar{f}_1)$ on the covariance graph. Its slope, $(\bar{f}_1 - \bar{f}_0)$, is a measure of forecast performance. This 'forecast slope' – later simply called 'slope' – is an indication of the forecaster's ability to discriminate between instances when the target outcome will and will not occur. The higher the slope, the better the forecaster is able to discriminate, and the better (lower) is the forecaster's Brier Score.

Yates (1982) showed that the forecast variance can be decomposed as follows:

$$s_f^2 = \text{min var } f + \text{scat } f ,$$

where

$$\text{min var } f = (\bar{f}_1 - \bar{f}_0)^2 \bar{d}(1 - \bar{d}) ,$$

$$\text{scat } f = (N_1, \text{var } f_1 + N_0 \text{ var } f_0)/N ,$$

and

N_1 = the number of occurrences of the target outcome,

N_0 = the number of non-occurrences of the target outcome,

N = the total number of probability forecasts ($N = N_1 + N_0$).

The component labeled 'min var f' is the minimum forecast variance that the forecaster can achieve while maintaining the level of discrimination ($\bar{f}_1 - \bar{f}_0$) between occasions when the target even does and does not occur. It is that part of the forecast variance that reflects the forecaster's ability to discriminate between outcomes. It is what the forecast variance would be if the conditional forecast variances, var f_1 and var f_0, were zero (i.e. if there were no scatter of forecasts about the conditional mean forecasts, $\bar{f}_1$ and $\bar{f}_0$). In contrast, scat f or 'scatter' is the weighted mean of the two conditional forecast variances. It is that part of the forecast variance that is not attributable to the forecaster's ability to discriminate between outcomes; it is excessive variance due primarily to the forecaster's reaction to non-predictive environmental cues. Ideally, this variance component would be zero.

The difference between the overall mean forecast and the mean outcome index ($\bar{f} - \bar{d}$) is referred to by Yates as the forecast 'bias'. The smaller the absolute value of the bias, the better 'calibrated-in-the-large' the forecaster is said to be [Yates and Curley (1985)].

We turn now to the description of covariance feedback.

3.2.4.2. Covariance feedback. Subjects in the covariance feedback group received (1) outcome feedback, (2) their Brier Scores and their individual component scores from the covariance decomposition [i.e. Brier Score = $\bar{d}(1 - \bar{d}) +$ min var f + scat $f + (\bar{f} - \bar{d})^2 - 2\,\mathrm{cov}(f, d)$; see Yates (1982)], (3) their individual covariance graphs, (4) the Brier Scores of the best and the average performer of the group, and (5) their ranking within the group as determined by the Brier Scores. Items (4) and (5) were used to motivate performance.

At the beginning of the second, third, and fourth sessions, the covariance decomposition was discussed and demonstrated. It was explained that for their forecasting problem the target outcome employed in the decomposition was 'home team wins the game'. [This is the same target outcome used by Yates (1982) and Yates and Curley (1985).] The meaning and significance of the components were explained using covariance graphs. The use of the covariance decomposition and covariance graphs to analyze forecasting performance was explained and demonstrated. Subjects were encouraged to construct probability forecasts that would maximize the slope of their covariance graph and minimize the scatter of their forecasts.

4. Results

To investigate the effects of the different forms of performance feedback on the external correspondence of subjects' probability forecasts, we analyzed the session-by-session performances in two ways. Like Lichtenstein and Fischhoff (1980), we analyzed the across-session performances within each treatment group, and like Sharp et al. (1988) we compared each group's performance with that of the control group. Forecasting performance was evaluated using the Brier Score and six performance measures derived from decompositions of the Brier Score: the calibration and resolution components of Murphy's decomposition, η^2 [the resolution measure suggested by Sharp et al. (1988) and described in Section 2], and bias, slope, and scatter from the covariance decomposition. In addition, a measure of overforecasting was employed. The mean of all probability forecasts minus the overall proportion correct ($\bar{f} - \bar{d}$) was used to measure overforecasting [Fischhoff and MacGregor (1982)]. A positive score indicates overforecasting, and a negative score reflects underforecasting.

Exhibits 2 through 5 present the means and standard deviations for the eight performance measures for each experimental group in each session, along with the proportion correct for each session as a measure of session difficulty. Statistically significant changes in group performance from one session to the next (as determined by paired-difference t-tests) are de-

noted by asterisks on the mean of the later session. Significant changes from the first to the last session are denoted by superscripts defined in the footnotes of the exhibits. In the case of the bias performance measure, the exhibits report mean bias scores (to preserve the information in the sign of the score), but the significance tests were conducted using absolute bias scores, the measure of calibration-in-the-large suggested by Yates and Curley (1985).

Ordinary-t-tests were used to compare the within-session performances of the feedback groups and the control group. The results of these tests are reported below but do not appear in the exhibits.

Finally, to determine whether forecasters manage their use of the probability scale differently when exposed to different forms of feedback and training, we describe any systematic differences in the probability values used by subjects of the different groups. The following subsections present the results of our analysis for each experimental group in turn. We begin with the control group.

4.1. Control group (Exhibit 2)

As expected, the provision of only outcome feedback was not sufficient to improve forecasting performance. For all but one measure, the performance of the control group either remained the same or deteriorated over the four sessions. Their Brier Scores, calibration, overforecasting, and scatter all deteriorated. Scatter deteriorated gradually over the four sessions and ended up significantly worse in session 4 than in session 1 ($p = 0.027$). No significant changes in resolution performance were observed using either Murphy's resolution component or η^2, and slope was essentially the same in the first and last sessions ($p = 0.33$). However, the control group did improve their 'calibration-in-the-large' (i.e. absolute bias). In other words, the average probability assigned to the target event 'home team wins' was closer to the actual proportion of home team wins in the later sessions.

This pattern of results can be explained in part by examining the changes in the group's usage of probability values across the four ses-

Exhibit 2
Means of performance measures for the control group.

	Session			
	1	2	3	4
Proportion correct	0.691	0.602	0.651	0.663
	(0.058)	(0.045)	(0.055)	(0.034)
Brier Score	0.208	0.248**[W]	0.239	0.231*[F]
	(0.037)	(0.031)	(0.038)	(0.030)
Calibration	0.020	0.042*[W]	0.040	0.043*[F]
	(0.011)	(0.022)	(0.030)	(0.030)
Overforecasting	0.034	0.131***[W]	0.102	0.112**[F]
	(0.080)	(0.066)	(0.102)	(0.098)
Resolution	0.023	0.030	0.026	0.034
	(0.019)	(0.020)	(0.012)	(0.016)
η^2	0.113	0.128	0.113	0.123
	(0.099)	(0.084)	(0.052)	(0.073)
Bias	−0.076	0.075	−0.013**[B]	−0.045*[L]
	(0.031)	(0.034)	(0.045)	(0.053)
Slope	0.238	0.176**[W]	0.193	0.227
	(0.118)	(0.087)	(0.054)	(0.087)
Scatter	0.057	0.068	0.075	0.083*[F]
	(0.021)	(0.024)	(0.038)	(0.042)

Standard deviations are given in parentheses.
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.0001$.
[W] Performance worse than previous session; [B] performance better than previous session.
[F] First session performance better than last session performance; [L] last session performance better than first session performance.

sions. The median number of different probabilities used by subjects was 6 in both sessions 1 and 4. However, subjects tended to shift their probability usage from lower values to higher values. In session 1, 39% of the forecasts were between 0.80 and 1.00 with 11% of these being 1.00s; in session 4 it increased to 56% with 20% being 1.00s. In addition, the control group was the only group of subjects that consistently used two-decimal probabilities in all four sessions. All other groups began with a mixture of one- and two-decimal probabilities, but by the third session every subject was using only one-decimal probabilities.

When not accompanied by increased knowledge, the observed shift in probability usage would tend to increase overforecasting and adversely affect calibration, scatter, and slope. This is consistent with the pattern revealed in Exhibit 2.

4.2. *Calibration-feedback group (Exhibit 3)*

As expected, the provision of calibration feedback resulted in improved forecasting performance. The calibration-feedback group's mean calibration score decreased significantly in the third session and maintained this higher performance level in the fourth session. As in Lichtenstein and Fischhoff's (1980) study of confidence judgments, improvement in calibration was not gradual, but occurred in one step. In their study, it occurred between sessions 1 and 2; in the present study, it occurred between sessions 2 and 3. Ten of the 15 subjects improved their calibration scores after the second session; no such improvement was observed for any of the other sessions. Similarly, overforecasting scores improved significantly in the third session and maintained that higher level in the fourth session. Ten of the 15 subjects achieved improved (i.e. reduced) overforecasting scores in session 3.

These findings were substantiated in comparisons with the control group. There were no significant differences in the mean calibration scores of the two groups in sessions 1 and 2, while in sessions 3 and 4 the calibration-feedback group's performance was superior ($p = 0.029$ for session 3; $p = 0.042$ for session 4). Similar results were observed for both overforecasting and scatter.

Exhibit 3
Means of performance measures for the calibration-feedback group.

	Session			
	1	2	3	4
Proportion correct	0.687	0.634	0.650	0.663
	(0.076)	(0.058)	(0.043)	(0.059)
Brier Score	0.209	0.237^{**W}	0.225	0.222
	(0.038)	(0.040)	(0.026)	(0.021)
Calibration	0.035	0.037	0.018^{*B}	0.023^{*L}
	(0.038)	(0.030)	(0.018)	(0.024)
Overforecasting	0.056	0.105^{*W}	0.045^{*B}	0.025^{*L}
	(0.084)	(0.073)	(0.061)	(0.062)
Resolution	0.024	0.029	0.021	0.018
	(0.014)	(0.019)	(0.019)	(0.022)
η^2	0.118	0.125	0.091	0.082
	(0.070)	(0.077)	(0.081)	(0.080)
Bias	−0.077	0.088	-0.013^{***B}	-0.066^{***W}
	(0.021)	(0.033)	(0.024)	(0.025)
Slope	0.254	0.196	0.165	0.160^{**F}
	(0.134)	(0.074)	(0.074)	(0.077)
Scatter	0.063	0.068	0.051	0.048
	(0.023)	(0.030)	(0.021)	(0.022)

Standard deviations are given in parentheses.
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.0001$.
[W] Performance worse than previous session; [B] performance better than previous session.
[F] First session performance better than last session performance; [L] last session performance better than first session performance.

As in Lichtenstein and Fischhoff's study, the improvements in calibration and overforecasting were not accompanied by significant worsening in resolution, whether measured by Murphy's resolution component or η^2. However, while not significant from session to session, slope deteriorated significantly between sessions 1 and 4 ($p = 0.004$). This suggests that the improvement in calibration and overforecasting may have been partly at the expense of discrimination.

Absolute bias (i.e. calibration-in-the-large) was not significantly different in session 4 than in session 1, although it improved significantly in session 3 and deteriorated significantly in session 4. No significant across-session trends were observed in scatter.

When the mean resolution, mean η^2, and mean slope scores were compared with those of the control group, the only significant differences occurred in session 4. The calibration-feedback group displayed the poorer performance on all three measures ($p = 0.026$, 0.045, and 0.034, respectively). This also suggests that the improved calibration of the calibration-feedback group came at the expense of discrimination. No significant differences in the mean Brier Score or mean absolute bias scores were observed between the two groups in any of the sessions (all p-values >0.05).

In contrast to the control group, subjects responded to calibration feedback and training by decreasing the number of different probabilities used (session 1 median was 6; session 4 median was 5) and by increasing their usage of lower probabilities and decreasing their usage of higher probabilities. In session 1, 44% of the forecasts were between 0.80 and 1.00 with 16% of all forecasts being 1.00s; in session 4, only 27% of the forecasts were between 0.80 and 1.00, with only 8% of all forecasts being 1.00s. Lichtenstein and Fischhoff (1980) observed a similar shift in probability usage in their calibration-feedback study. This change in probability usage would tend to improve a forecaster's overforecasting, but would also tend to reduce the slope of the forecaster's covariance graph. This is consistent with the results described above.

4.3. Resolution-feedback group (Exhibit 4)

Contrary to our expectations and to the results obtained by Sharp et al. (1988), the resolution feedback and training did not affect the group's resolution performance. Neither the res-

Exhibit 4
Means of performance measures for the resolution-feedback group.

	Session			
	1	2	3	4
Proportion correct	0.660	0.635	0.600	0.651
	(0.042)	(0.062)	(0.052)	(0.060)
Brier Score	0.218	0.229	0.251^{*W}	0.231^{*B}
	(0.026)	(0.032)	(0.039)	(0.027)
Calibration	0.023	0.030	0.041^{*W}	0.031
	(0.016)	(0.026)	(0.025)	(0.023)
Overforecasting	0.025	0.050	0.124^{**W}	0.060^{**B}
	(0.072)	(0.089)	(0.091)	(0.091)
Resolution	0.028	0.030	0.031	0.023
	(0.018)	(0.020)	(0.031)	(0.029)
η^2	0.126	0.131	0.128	0.105
	(0.074)	(0.079)	(0.105)	(0.103)
Bias	−0.054	0.063	-0.013^{*B}	-0.074^{***W}
	(0.057)	(0.040)	(0.029)	(0.030)
Slope	0.174	0.167	0.158	0.191
	(0.082)	(0.076)	(0.078)	(0.087)
Scatter	0.045	0.052	0.073^{*W}	0.067
	(0.024)	(0.028)	(0.042)	(0.042)

Standard deviations are given in parentheses.
* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.
[W] Performance worse than previous session; [B] performance better than previous session.

olution score nor η^2 changed significantly over the four sessions of the experiment. Although some of the performance measures varied significantly in sessions 3 and 4, none of the measures of session 4 differed significantly from session 1.

No significant differences in any of the performance measures were observed for the resolution-feedback group and the control group in any of the sessions (all p-values >0.05). Both groups showed basically the same trends across sessions for each of the eight performance measures. Any session-to-session differences appear to be due to the difficulty levels of the sessions as measured by the proportion correct. The control group had the most difficulty with the forecasts of session 2 (see Exhibit 2); the resolution-feedback group had the most difficulty with session 3 (see Exhibit 4).

The resolution-feedback group increased its usage of 0.5 and 1.0 probabilities. In fact, this group steadily increased its usage of 0.5 and 1.0 probabilities across the four experimental sessions. In session 1, 21% of the forecasts were 0.5 and 5% were 1.0; in session 4, 32% were 0.5 and 18% were 1.0. In addition, all subjects resorted to using only a few different probabilities in their attempts to improve their resolution scores. In session 1, the median number of different probabilities used was 6, but this dropped to 3 in session 4. In fact, by the third session, most subjects used only 0.5, 1.0, and a 'middle-of-the-road' probability. In contrast, all control group subjects used at least five different probabilities in session 4. Also, the control group increased its use of higher probabilities and decreased its use of lower probabilities over the course of the study. Both groups, however, became heavy users of 1.0 in the later sessions.

Such extensive use of extreme probabilities would typically increase the scatter of the forecasts. Further, extensive use of 1.0 without a significant increase in knowledge would hurt both calibration and overforecasting. These patterns were observed for both the resolution-feedback group and the control group.

4.4. Covariance-feedback group (Exhibit 5)

Contrary to our expectations, the covariance feedback we provided was ineffective. Absolute bias (i.e. calibration-in-the-large) was the only measure on which the covariance-feedback

Exhibit 5
Means of performance measures for the covariance-feedback group.

	Session			
	1	2	3	4
Proportion correct	0.682	0.645	0.644	0.642
	(0.092)	(0.047)	(0.046)	(0.043)
Brier Score	0.213	0.238^{*W}	0.233	0.243
	(0.031)	(0.038)	(0.025)	(0.039)
Calibration	0.028	0.042	0.034	0.043^{*F}
	(0.017)	(0.028)	(0.025)	(0.031)
Overforecasting	0.005	0.102^{**W}	0.087	0.102^{*F}
	(0.087)	(0.095)	(0.075)	(0.089)
Resolution	0.023	0.030	0.028	0.028
	(0.011)	(0.018)	(0.014)	(0.011)
η^2	0.112	0.134	0.124	0.122
	(0.044)	(0.079)	(0.061)	(0.050)
Bias	−0.081	0.067^{*B}	-0.011^{***B}	$-0.068^{***W\ *L}$
	(0.019)	(0.020)	(0.038)	(0.027)
Slope	0.200	0.206	0.191	0.184
	(0.108)	(0.066)	(0.051)	(0.047)
Scatter	0.048	0.077^{**W}	0.069	0.079
	(0.027)	(0.046)	(0.027)	(0.039)

Standard deviations are given in parentheses.
* $p<0.05$; ** $p<0.01$; *** $p<0.0001$.
[W] Performance worse than previous session; [B] performance better than previous session.
[F] First session performance better than last session performance; [L] last session performance better than first session performance.

group showed significant improvement. Improvements were observed in both sessions 2 and 3. While the group's performance deteriorated significantly in session 4, their session 4 performance was significantly better than in session 1 ($p = 0.034$). The Brier Score, overforecasting, scatter, and calibration all deteriorated. For calibration, the deterioration was gradual; the others deteriorated in one step, between sessions 1 and 2. No significant changes in resolution, η^2, or slope performance were observed.

As was the case for the resolution-feedback group, no significant differences in any of the performance measures were observed for the covariance-feedback group and the control group in any of the sessions (all p-values >0.05). Also, no significant differences were observed between the performances of the covariance-feedback group and the resolution-feedback group (all p-values >0.05). Thus, the provision of covariance feedback and training apparently were no more effective than either resolution feedback and training or simple outcome feedback.

Like the control group, the median number of different probabilities used was 6 in both sessions 1 and 4. Also like the control group, the covariance-feedback group decreased its use of 0.5 probabilities over the course of the study (from 25% of their forecasts in session 1 to 14% in session 4) and increased its use of 1.0s (from 9% to 21%). Nearly all of the covariance-feedback subjects (13 of the 16 subjects) reported being overwhelmed by the feedback. Most indicated that because there was too much information to deal with, they focused primarily on trying to improve the slope of their covariance graphs. The above-noted shift in probability usage is consistent with such efforts.

5. Discussion

For confidence judgments in general-knowledge tasks, Lichtenstein and Fischhoff (1980) found that calibration feedback and training improved both the subjects' calibration and overconfidence performances. The present study generalizes those results to probability forecasting tasks. For the calibration-feedback group, we observed a significant improvement in calibration and overforecasting relative to the control group. Furthermore, as in Lichtenstein and Fischhoff's study, virtually all of the improvement in calibration occurred in one step. However, in their study the improvement occurred in the second session, while in ours it occurred in the third session. While this disparity could be a result of the differences in the feedback and training of the two studies, we believe it is due to differences in the tasks.

In particular, there are fundamental differences in what can be learned from feedback in the two tasks. In general-knowledge tasks, subjects respond to a series of almanac-type questions drawn from different subject domains. The information provided by feedback – whether it be outcome feedback, performance feedback, or both – bears only on the subject's probability usage. It helps the subject assess the appropriateness of her expressions of confidence. Answers to unrelated almanac questions (i.e. outcome feedback) will not help the subject to answer future questions. In a series of forecasting tasks, however, particularly when the events being predicted are related as in the present study, feedback may inform the subject about not only the appropriateness of her probability usage, but about external reality as well (e.g. the predictability of the events in question). Thus, feedback may lend support not only to the judgmental process the subject uses to assign a number to her predictive belief, but also to the reasoning process [Smith et al. (1991)] that generated the belief. In forecasting tasks, it may simply take more experience with the task and more than one round of feedback to realize the benefits of feedback. [For other differences between forecasting and general-knowledge tasks, see Wright and Ayton (1986).]

Based on the results of the control group, outcome feedback was not sufficient to improve calibration and overforecasting. In fact, both deteriorated in this study. On all performance measures except absolute bias, the control group's performance in session 4 was either unchanged or worse than in session 1. This general inability of outcome feedback to improve probability judgments is consistent with previous findings [e.g. Fischer (1982)].

Our resolution and covariance feedback and training turned out to be no more effective than

outcome feedback. There were no statistically significant differences between the performances of these groups and the control group for any of the eight performance measures in any of the four sessions. We believe that the failure of the resolution feedback was due primarily to the fact that it did nothing to improve the substance of the subjects' knowledge about the events being forecasted. The training was designed to help the subjects understand the resolution concept, to motivate them to improve their resolution, to discourage them from using high probabilities when they had little information to go on, and to help them sort their forecasts into categories based on the extent of their knowledge of the events in question. However, it provided no new information to the subjects about the games or teams for which they were asked to make predictions; in other words, it provided no environmental feedback.

Even though we encouraged the use of meaningful probability values, resolution-group subjects apparently focused more strongly on distinct forecast categories than on the probabilities associated with the categories. This was evidenced by their increased use of 0.5s and 1.0s over the course of the study and by their post-experiment interviews. In addition, subjects' use of only three or four different probability values following the receipt of feedback may have been due to our use of only four knowledge categories in the feedback and training. Future studies should permit subjects to sort the events to be forecast into as many different knowledge categories as they care to use.

We believe that the failure of the covariance-feedback group to outperform the control group was due in part to the amount of feedback provided to subjects in each session. Subjects received their Brier Scores, four component scores from the covariance decomposition, a covariance graph, outcome feedback, information on the Brier Scores of others in their group, and the rank of their Brier Score within the group. In addition, we encouraged the subjects to minimize their Brier Score by maximizing their slope and minimizing their scatter scores. Having to attend to so much information and to multiple objectives probably resulted in cognitive overload. In post-experiment interviews, three subjects indicated some 'confusion' over what to do with the feedback, and 13 subjects complained of 'too much information in the feedback'. None of the subjects in any of the other groups made such comments. By comparison, the performance feedback and training provided to the other groups was much more focused. Future studies should either reduce the amount of feedback (for example, to just slope and scatter) or increase the number of feedback sessions to give subjects sufficient time to understand and exploit the session-to-session variation in the various feedback components.

The differential effects of feedback are reflected in the probability values the subjects chose to use. All groups that received performance feedback (i.e. all groups but the control) shifted from using two-digit probabilities to one-digit probabilities. In addition, both the calibration and resolution groups used fewer different probabilities in later sessions. These results suggest that the provision of focused performance feedback and training (i.e. that received by the calibration and resolution feedback groups) led subjects to reduce the number of probability categories to which they attended in order to better manage their forecasts relative to the performance incentives.

6. Conclusion

We began this paper by describing the four types of feedback that are relevant in judgmental forecasting tasks: outcome, performance, process, and environmental feedback. We end the paper by describing what has been learned about these feedback types and what questions remain unanswered.

This study has confirmed earlier work finding that forecasters need more than just outcome feedback to improve the accuracy of their forecasts. The information content of outcome feedback apparently is not sufficient to either increase forecasters' knowledge of the event in question or to help forecasters assign better probability labels to their forecasts. Something more is needed.

We found that the additional information provided by focused, personalized performance feedback did help improve forecast accuracy. In particular, calibration feedback and training was

shown to improve forecasters' abilities to assign meaningful probability labels to their forecasts (i.e. to improve their calibration and overforecasting). Such improvement is critically important to forecast users. The better calibrated the forecaster, the more her probability forecasts are like relative frequencies, and the easier they are to interpret and use. For example, having received a 0.8 probability of the stock market rising from a well-calibrated securities analyst, the forecast user need not be concerned with how to interpret the probability or how much to adjust the forecast to compensate for overforecasting. The user knows that 80% of the time when a 0.8 forecast is issued, the event will occur.

Several questions concerning calibration feedback remain unanswered and await future research: Will calibration performance deteriorate if calibration feedback is cut off? Will the effects of calibration feedback and training for one forecasting task transfer to another forecasting task? Can further improvement in calibration be realized through other types of feedback? The results of the present study are strong enough, however, that we recommend that practitioners not wait for the answers to these questions to begin exploiting calibration feedback. It should be employed both in training probability forecasters and as part of a program of periodic, personalized performance feedback.

Our results suggest that improvement in forecasters' discrimination skills (i.e. resolution) requires more than comprehension of the resolution concept and related performance feedback; it requires better use of the forecaster's existing knowledge of the event in question or an increase in that knowledge. The former could be accomplished through process feedback; the latter through environmental feedback. We believe that process and environmental feedback represent significant opportunities for improving the accuracy of probability forecasters. They should figure prominently in future studies of judgmental forecasting.

Acknowledgements

The authors wish to acknowledge the helpful comments and suggestions of Shawn P. Curley. This research was supported by a University of Minnesota Doctoral Dissertation Special Grant and by the Research Committee of the Carlson School of Management, University of Minnesota.

References

Balzer, W.K., M.E. Doherty and R. O'Connor, Jr., 1989, "Effects of cognitive feedback on performance", *Psychological Bulletin*, 106, 410–433.

Beach, B.H., 1975, "Expert judgment about uncertainty: Bayesian decision making in realistic settings", *Organizational Behavior and Human Performance*, 14, 10–59.

Brehmer, B., 1980, "In one word: Not from experience", *Acta Psychologica*, 45, 223–241.

Brier, G.W., 1950, "Verification of forecasts expressed in terms of probability", *Monthly Weather Review*, 78, 1–3.

Fischer, G.W., 1982, "Scoring-rule feedback and the overconfidence syndrome in subjective probability forecasting", *Organizational Behavior and Human Performance*, 29, 352–369.

Fischhoff, B. and D. MacGregor, 1982, "Subjective confidence in forecasts", *Journal of Forecasting*, 1, 155–172.

Fischhoff, B., P. Slovic and S. Lichtenstein, 1977, "The appropriateness of extreme confidence", *Journal of Experimental Psychology: Human Perception and Performance*, 3, 552–564.

Friedman, D., 1983, "Effective scoring rules for probabilistic forecasts", *Management Science*, 29, 447–454.

Kidd, J.B., 1973, "Scoring rules for subjective assessments", A paper written for the Annual Conference of the Operational Research Society, Torbay, England.

Lichtenstein, S. and B. Fischhoff, 1980, "Training for calibration", *Organizational Behavior and Human Performance*, 26, 149–171.

Lichtenstein, S., B. Fischhoff and L.D. Phillips, 1982, "Calibration of probabilities: The state of the art to 1980", in: D. Kahneman, P. Slovic and A. Tversky, eds., *Judgment Under Uncertainty: Heuristics and Biases* (Cambridge University Press, Cambridge).

Murphy, A.H., 1973, "A new vector partition of the probability score", *Journal of Applied Meteorology*, 12, 595–600.

Murphy, A.H. and H. Daan, 1984, "Impacts of feedback and experience on the quality of subjective probability forecasts: Comparison of results from the first and second years of the Zierikzee experiment", *Monthly Weather Review*, 112, 413–423.

Murphy, A.H., W. Hsu, R.L. Winkler and D.S. Wilks, 1985, "The use of probabilities in subjective quantitative precipitation forecasts: Some experimental results", *Monthly Weather Review*, 113, 2075–2089.

Ronis, D.L. and J.F. Yates, 1987, "Components of probability judgment accuracy: Individual consistency and effects of subject matter and assessment method", *Organizational Behavior and Human Decision Processes*, 40, 193–218.

Sharp, G.L., B.L. Cutler and S.D. Penrod, 1988, "Performance feedback improves the resolution of confidence judgments", *Organizational Behavior and Human Decision Processes*, 42, 271–283.

Smith, G.F., P.G. Benson and S.P. Curley, 1991, "Belief, knowledge, and uncertainty: A cognitive perspective on subjective probability", *Organizational Behavior and Human Decision Processes*, 48, 291–321.

Staël von Holstein, C.-A.S., 1972, "Probabilistic forecasting: an experiment related to the stock market", *Organizational Behavior and Human Performance*, 8, 139–158.

Winkler, R.L., 1969, "Scoring rules and the evaluation of probability assessors", *Journal of the American Statistical Association*, 64, 1073–1078.

Wright, G. and P. Ayton, 1986, "Subjective confidence in forecasts: a response to Fischhoff and MacGregor", *Journal of Forecasting*, 5, 117–123.

Yates, J.F., 1982, "External correspondence: Decompositions of the mean probability score", *Organizational Behavior and Human Performance*, 30, 132–156.

Yates, J.F. and S.P. Curley, 1985, "Conditional distribution analyses of probabilistic forecasts", *Journal of Forecasting*, 4, 61–73.

Biographies: P. George BENSON is an Associate Professor of Decision Sciences and Operations Management at the Curtis L. Carlson School of Management, University of Minnesota. He received a B.S. in Mathematics from Bucknell University and a Ph.D. in Decision Sciences from the University of Florida. Professor Benson's current research interests include probability forecasting, belief assessment, decision analysis, quality management, and methods for continuous improvement. He has published articles in numerous journals including *Management Science*, the *Journal of Forecasting*, *Organizational Behavior and Human Decision Processes*, *Decision Sciences*, the *Journal of Finance*, and the *Journal of Quality Technology*.

Dilek ÖNKAL is an Assistant Professor of Decision Sciences at Bilkent University, Turkey. She received a Ph.D. in Decision Sciences from the University of Minnesota. Her research interests include probability forecasting and reliability of subjective probabilities. She has published in the *International Forum for Information and Documentation*.